BAB I PENGANTAR RAG DAN EMBEDDING

Retrieval-Augmented Generation (RAG) adalah teknik dalam kecerdasan buatan yang menggabungkan proses pencarian informasi (retrieval) dengan pembuatan teks (generation). Sistem RAG dirancang untuk mengatasi keterbatasan Large Language Model (LLM) seperti halusinasi dan ketidaktahuan terhadap data privat.

Alur RAG terbagi menjadi dua tahap utama: Indexing (offline) dan Query (saat pertanyaan datang). Pada tahap Indexing, dokumen dibaca, dipotong menjadi bagian kecil (chunking), diubah menjadi vektor angka (embedding), lalu disimpan ke dalam vector database seperti ChromaDB.

BAB II CHUNKING DAN STRATEGI PEMOTONGAN TEKS

Chunking adalah proses memotong dokumen besar menjadi potongan-potongan kecil agar lebih mudah diproses oleh model embedding. Ukuran chunk yang terlalu kecil dapat menyebabkan hilangnya konteks, sementara ukuran yang terlalu besar membuat relevansi informasi menjadi kabur.

Penggunaan overlap (tumpang tindih) antar chunk sangat penting untuk memastikan informasi penting di perbatasan kalimat tidak terpotong atau terbelah. Ukuran umum yang sering digunakan berkisar antara 500 hingga 1500 karakter dengan overlap sekitar 10 hingga 20 persen.

BAB III EVALUASI PENCARIAN DENGAN HIT@K

Evaluasi retrieval sangat penting untuk mengukur apakah sistem berhasil menemukan dokumen yang relevan di peringkat teratas. Metrik Hit@k (seperti Hit@1, Hit@3, dan Hit@5) digunakan untuk mengecek apakah jawaban atau kata kunci yang benar berada di dalam k hasil teratas dari pencarian vektor.